Does Markdown Increase LLM Bot Traffic? A Randomized Controlled Experiment

Brandon Punturo

February 2026

Abstract

We conducted a randomized controlled experiment across 381 pages from six websites to test whether serving Markdown (vs. HTML) to LLM crawlers increases bot visit volume. Pages were randomly assigned to treatment (Markdown) or control (HTML) groups, and bot traffic was measured over 20 days (January 19 – February 8, 2026) using Profound's Agent Analytics product. We found no statistically significant difference in overall bot traffic: the median visits per page were 6.0 (control) vs. 7.0 (treatment), and the CUPED-adjusted mean lift was +13.9% (95% CI: [−11.2%, +46.8%]) [Deng et al., 2013]. ChatGPT-User showed a directional advantage for Markdown (+8.5% to +20% depending on specification), but the signal was not statistically significant. We conclude that Markdown does not produce "slam dunk" gains in bot traffic, and any effect, if present, is likely small.

Contents

1. Introduction & Motivation

Over the past several months, claims have proliferated across SEO and AI optimization communities that serving Markdown to LLM crawlers increases visibility in AI-powered answer engines like ChatGPT, Perplexity, and Claude. The hypothesis is intuitive: Markdown is a cleaner, more structured format than HTML, which may make it easier for large language models to parse, extract, and incorporate content into their training data or retrieval systems.

However, most evidence supporting this claim has been anecdotal—individual case studies, before-after comparisons, or correlational observations. No randomized controlled experiment has been published testing the causal effect of content format on bot traffic. This is the gap our study addresses.

1.1 Why This Matters

LLM crawlers are becoming a significant source of organic traffic for content publishers. AI-sourced web traffic surged by 527% year-over-year between January and May 2025, growing from approximately 17,000 to over 107,000 sessions across analyzed properties [Search Engine Land, 2025b]. Cloudflare's network alone recorded approximately 50 billion crawler requests per day from AI bots by late 2025 [Thunderbit, 2026], and the share of web traffic attributable to AI-oriented bots reached 4.2% of all HTML page requests that year [Cloudflare, 2025a]. As AI-powered search tools gain market share, understanding how to optimize for bot crawl behavior—and ultimately, for inclusion in AI-generated answers—has become a strategic priority [Previsible, 2026, Akamai, 2025].

If serving Markdown drives measurably higher bot traffic, it would represent a low-hanging fruit for content optimization. Conversely, if the effect is negligible, publishers can avoid investing engineering resources in a strategy that does not move the needle.

1.2 Our Contribution

This paper presents a randomized controlled experiment testing the effect of serving Markdown (vs. HTML) to LLM crawlers on bot visit volume. We use page-level randomization across six real-world websites, variance-reduction techniques (CUPED) [Deng et al., 2013], and rigorous statistical testing to estimate the causal effect.

Our findings align with skepticism expressed by practitioners at major search platforms. Google's John Mueller stated that "LLMs have trained on—read & parsed—normal web pages since the beginning. It seems a given that they have no problems dealing with HTML" [Mueller, 2025]. Microsoft's Fabrice Canel echoed this position, noting that "AI makes us great at understanding web pages" and questioning whether separate bot-only content formats add any meaningful signal [Goodwin, 2026]. A parallel observational study by OtterlyAI [OtterlyAI, 2026]—examining `llms.txt` file adoption rather than Markdown serving—similarly found no positive correlation between format signals and AI crawler activity, consistent with our null result.

2. Experimental Design

2.1 Sample

We partnered with six websites spanning diverse content verticals. In total, 380 pages were included in the experiment. Pages were randomized at the individual level, stratified by website, into two arms:

- **Control (HTML):** 189 pages
- **Treatment (Markdown):** 191 pages (after excluding 1 outlier)

Table 1: Page distribution across websites by treatment arm.

Website	Control Pages	Treatment Pages
Website 1	30	30
Website 2	34	34 (1 outlier removed)
Website 3	40	40
Website 4	33	34
Website 5	38	38
Website 6	14	15
Total	**189**	**191**

2.2 Treatment

The treatment consisted of serving clean Markdown versions of page content to detected LLM bots, while serving standard HTML to all human visitors and non-LLM crawlers. Implementation was handled via middleware, which intercepted requests, detected bot user agents, and served the appropriate format dynamically.

2.3 Bot Detection & Classification

We classified requests as "LLM bot traffic" if they matched known user agent patterns for:

- **OpenAI:** `ChatGPT-User`, `OAI-SearchBot`, `GPTBot`
- **Anthropic:** `ClaudeBot`
- **Perplexity:** `PerplexityBot`
- **Meta:** `Meta-ExternalAgent`, `FacebookBot`
- **DuckDuckGo:** `DuckAssistBot`

Traditional search engine crawlers (Googlebot, Bingbot) were excluded, as it is difficult to distinguish LLM-related activity (e.g., indexing for AI Overviews) from standard search indexing. The goal was to isolate bot traffic clearly attributable to AI-powered tools.

2.4 Primary Metric

The primary outcome was total bot visits per page during the experiment period. A "visit" was defined as a unique request from an LLM bot to a tracked page, deduplicated at 1-minute granularity (i.e., multiple requests from the same bot to the same page within a 1-minute window counted as a single visit).

2.5 Duration & Timeline

- **Pre-period:** January 2, 2026 – January 12, 2026 (used for CUPED covariate)
- **Experiment period:** January 19, 2026 – February 8, 2026 (20 days)

3. Data Pipeline

3.1 Source Data

Bot traffic data was sourced from Agent Analytics, Profound's product for measuring LLM bot activity.

3.2 Deduplication

Raw event logs contained 8,151 bot requests. After deduplication at 1-minute granularity, 957 duplicate events were removed, leaving 7,194 unique visits for analysis.

3.3 Outlier Handling

One page received 345 visits from Meta/Facebook bots, roughly 10× more than any other page. This was classified as an outlier and excluded from analysis. After exclusion:

- Control: 189 pages
- Treatment: 191 pages

All subsequent analyses use the outlier-excluded dataset unless otherwise noted.

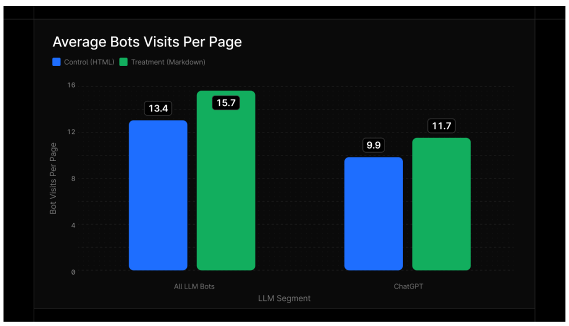
4. Results

4.1 Primary Analysis: Aggregate Metrics

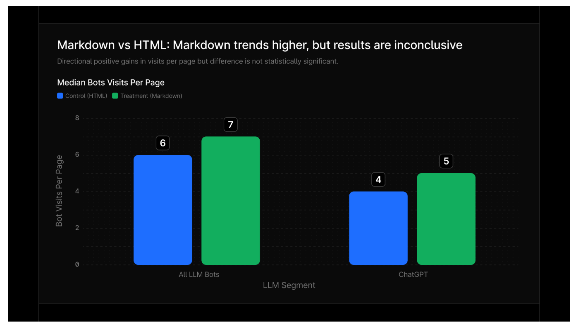
Table 2 summarizes key metrics across treatment arms.

Table 2: Aggregate bot visit metrics by treatment arm.

Metric	Control (HTML)	Treatment (Markdown)
Median visits per page	6.0	7.0
Mean visits per page	13.5	15.7
Total bot visits	2,542	2,992
Pages	189	191



(a) Mean visits per page



(b) Median visits per page

Figure 1: Median and mean bot visits per page by treatment arm. The Markdown group shows a slight directional advantage, but statistical tests indicate this difference is consistent with random noise.

The raw numbers suggest a directional advantage for Markdown:

- Median: +1 visit per page (+16.7%)
- Mean: +2.2 visits per page (+16.5%)
- Total: +450 visits (+17.7%)

However, raw comparisons do not account for pre-existing differences between pages or variance in baseline traffic.

4.2 Statistical Tests

We conducted three primary statistical tests.

4.2.1 Mann-Whitney U Test (Non-Parametric)

The Mann-Whitney U test compares the probability that a randomly selected page from the treatment group has more visits than a randomly selected page from the control group.

Table 3: Mann-Whitney U test results.

Test	**Estimate**	***p*-value**
Mann-Whitney U (all bots)	$P(\text{treat} > \text{ctrl}) = 0.515$	0.62
Mann-Whitney U (ChatGPT only)	$P(\text{treat} > \text{ctrl}) = 0.526$	0.38

A probability of 0.515 is only slightly above 0.50 (which would indicate no difference). The p-value of 0.62 indicates we cannot reject the null hypothesis of no effect.

4.2.2 CUPED-Adjusted Mean (Variance Reduction)

CUPED (Controlled-experiment Using Pre-Experiment Data) is a variance-reduction technique that adjusts treatment outcomes based on pre-experiment traffic [Deng et al., 2013]. The estimator is:

$$\hat{Y}_{\text{adj}} = Y - \theta\left(X - \bar{X}\right) \tag{1}$$

where:

- Y = visits during experiment period
- X = visits during pre-period
- $\bar{X}$ = mean of pre-period visits
- $\theta = \text{Cov}(Y, X)/\text{Var}(X)$ (regression coefficient)

This reduces noise from pages with inherently high or low traffic, improving statistical power.

Table 4: CUPED-adjusted lift estimates.

Covariate Window	**Overall Lift**	**95% CI**	**ChatGPT Lift**	**95% CI**
Raw (no CUPED)	+16.5%	—	+18.4%	—
Jan 2 – Jan 12	+13.9%	[−11.2%, +46.8%]	+8.5%	[−15.0%, +38.8%]

Why ChatGPT's CUPED adjustment is larger. The raw data show ChatGPT lift (+18.4%) exceeding overall lift (+16.5%), but CUPED reverses this: ChatGPT (+8.5%) falls below overall (+13.9%). This occurs because: (1) the treatment group had a larger pre-period imbalance for ChatGPT traffic (+8%) than for overall bot traffic (+2%)—pure randomization noise with $\sim$190 pages per arm; and (2) ChatGPT's regression coefficient was higher ($\theta = 2.71$ vs. 2.23 overall), meaning pre-period traffic was more predictive and CUPED applied a stronger correction (56% variance reduction vs. 39%). In essence, the treatment group got "lucky" in the randomization draw for ChatGPT visits, and CUPED correctly adjusts for this pre-existing advantage.

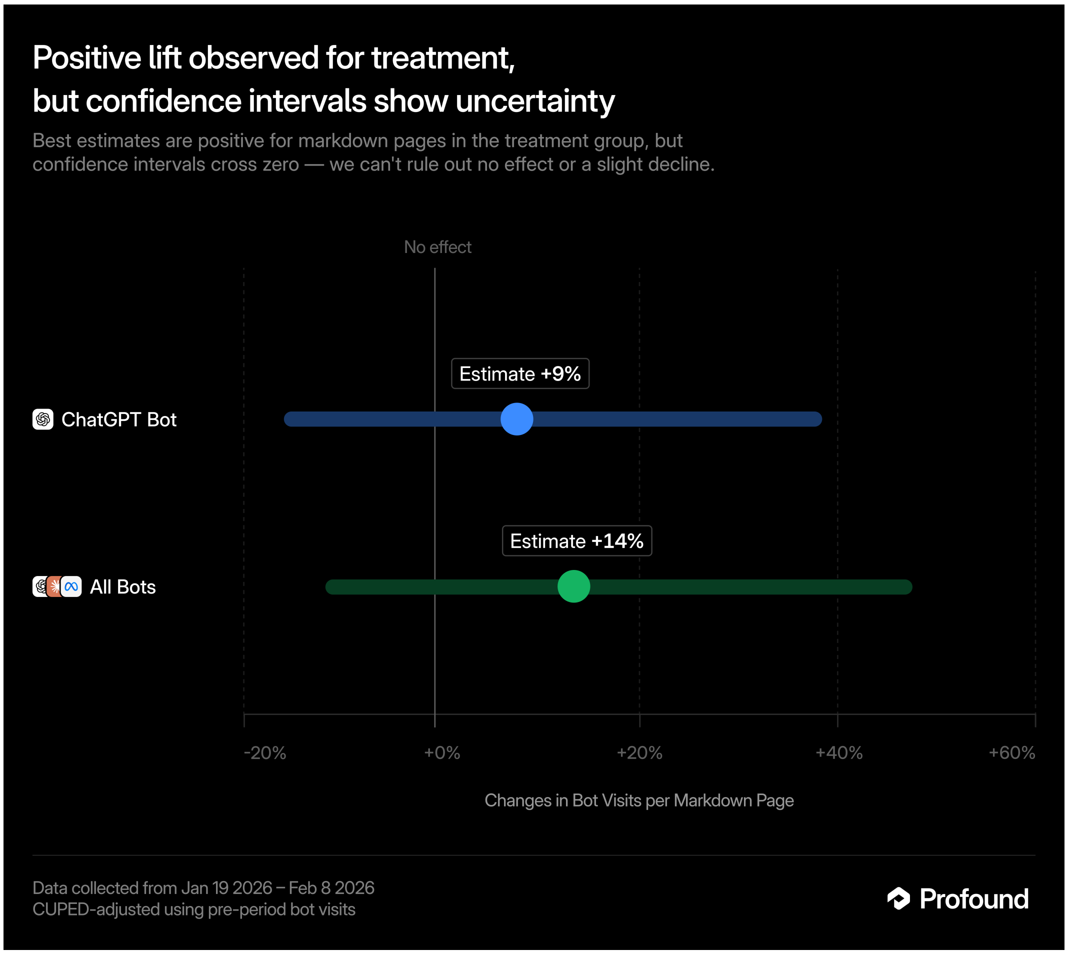


Figure 2: Forest plot of CUPED-adjusted lift estimates. All confidence intervals span zero, indicating no statistically significant effect.

All confidence intervals include zero, meaning the true effect could plausibly be negative, zero, or positive. None of the specifications reach statistical significance. CUPED achieved variance reduction of $\sim$40%, substantially reducing noise compared to raw comparisons.

Pre-period window justification (Jan 2–12). We selected January 2–12 as the CUPED covariate window for two reasons. First, December 25–January 1 exhibits holiday-period traffic patterns unrepresentative of steady-state conditions and would add noise to the covariate. Second, January 13–18 overlaps with the implementation period during which the experiment's bot-detection infrastructure was being deployed and tested; traffic patterns during this window may have been influenced by the instrumentation rollout itself.

4.2.3 Median Permutation Test

To test whether the +1 median difference was meaningful, we ran a permutation test: we randomly shuffled treatment labels 10,000 times and calculated the median difference each time. The p-value is the fraction of permutations where the shuffled difference was as large or larger than the observed difference.

Table 5: Median permutation test results.

Test	**Observed Diff.**	**Simulated 95% CI**	***p*-value**
Median permutation (all bots)	+1.0 visits	[−2.0, +3.0]	1.000
Median permutation (ChatGPT only)	+1.0 visits	[−1.0, +2.0]	0.689

A p-value of 1.000 means the observed difference is completely within the range of random noise. In other words, even if Markdown had no effect, we would expect to see a +1 median difference purely by chance nearly 100% of the time.

4.3 Bot-Level Breakdown

Not all bots behaved the same way. Table 6 shows total visits and lift by bot family.

Table 6: Bot visits and winsorized lift by bot family.

Bot	**Control**	**Treatment**	**Winsorized Lift**
ChatGPT-User	1,870	2,238	+20%
Meta/Facebook	508	585	0%
OAI-SearchBot	95	99	−1%
ClaudeBot	47	43	−4%

ChatGPT-User was the only bot with a consistently positive signal. This bot represents user-triggered searches via ChatGPT's web browsing feature and accounted for 73% of all bot traffic in the experiment.

Table 7: Traffic composition by bot family.

Bot Name	Percentage of Traffic
ChatGPT-User	73%
Meta/Facebook	20%
OAI-SearchBot	4%
ClaudeBot	2%
GPTBot	1%

ChatGPT-User dominates traffic, with Meta/Facebook a distant second.

Other bots (Meta, OAI-SearchBot, ClaudeBot, GPTBot) showed flat or slightly negative effects. This suggests that if Markdown has any effect, it may be bot-specific rather than universal.

4.4 Distribution Analysis

Table 8: Distribution of bot visits per page by percentile.

Percentile	Control	Treatment	Difference
Min	0	0	0
p10	1	0	−1
p25	3	2	−0
p50 (Median)	6	7	+1
p75	16	18	+2
p90	27	38	+11
p95	47	58	+11
p99	128	114	−13
Max	168	160	−8

The distribution shows substantial overlap at lower percentiles, with the treatment group exhibiting a rightward shift in the upper tail (p75–p95)—consistent with the higher mean but similar median.

4.4.1 Percentile Analysis (ChatGPT Only)

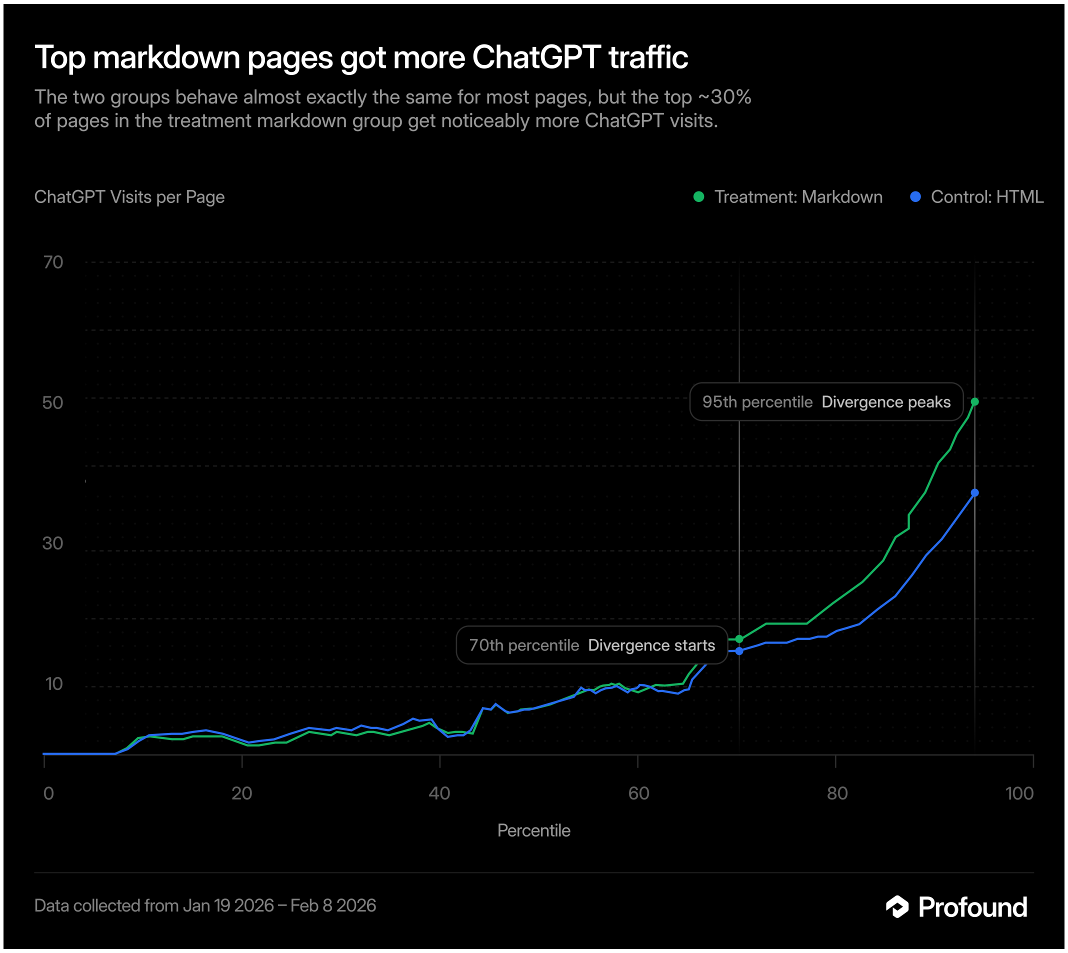


Figure 3: Percentile comparison for ChatGPT-User visits. Below the median, the two groups are nearly identical. Above the 60th percentile, the treatment group pulls ahead.

The Markdown advantage is concentrated in pages that already receive high bot traffic. Below the median (pages with ≤ 4 visits), there is no difference. The divergence begins around the 60th percentile and widens at the top. This suggests that Markdown may amplify crawl activity on pages already being visited frequently, rather than driving new bot discovery.

4.5 Website-Level Heterogeneity

Table 9: Lift by website, including pre-period imbalance.

Website	Pre-Period Lift	Experiment Lift	ChatGPT Lift
Website 1	−10%	+21%	+27%
Website 2	+10%	+45%	+32%
Website 3	−8%	−8%	−10%
Website 4	−21%	+10%	+57%
Website 5	+61%	+27%	+30%
Website 6	+2%	−18%	−18%

Sites with the largest apparent experiment effects also had notable pre-period imbalances, suggesting some of the observed "lift" may be due to noise in randomization rather than a true treatment effect. For example:

- Website 5 showed +27% lift during the experiment, but the treatment group already had +61% more traffic in the pre-period.
- Website 3 showed −8% lift during the experiment, matching the −8% imbalance in the pre-period.

This pattern is consistent with random variation rather than a causal effect.

4.6 Time Series Analysis

Table 10: Average daily lift by time period.

Period	Days	Avg. Daily Lift
Pre-Period	10	2%
Implementation Period (Jan 13–18)	6	12%
Experiment	21	18%

Table 11: Average daily lift by time period (with context).

Period	Days	Avg. Lift
Pre-Period (Jan 2–12)	10	+4%
Implementation (Jan 13–18)	6	+10%
Experiment (Jan 19–Feb 8)	21	+15%

Tables 10 and 11 summarize the daily lift patterns across different periods. The time series shows:

- **High day-to-day volatility:** Daily lift ranges from −40% to +60%, even during the pre-period (when both groups received the same HTML).
- **No temporal trend:** There is no evidence that the Markdown effect grew or shrank over the course of the experiment.
- **Pre-period noise matches experiment noise:** The magnitude of fluctuation during the pre-period is similar to the experiment period, suggesting the noise is inherent to bot behavior rather than treatment-related.

5. Power Analysis

Power calculations are reported for the mean-based estimators. The Mann-Whitney and median permutation tests have comparable or coarser detection thresholds at this sample size.

5.1 Minimum Detectable Effect (MDE)

Table 12: Minimum detectable effect at 80% power, $\alpha = 0.05$.

Specification	**MDE (80% power)**
Raw (no CUPED)	$\sim 53\%$
CUPED-adjusted	$\sim 41\%$

With 189–191 pages per arm, this experiment could reliably detect an effect of $\sim 41\%$ or larger (with CUPED). Effects below 41% would often be missed due to insufficient sample size.

5.2 What We Can Conclude

While the experiment was underpowered for moderate effects, it was adequately powered to rule out large effects. Specifically:

- If Markdown drove a $> 41\%$ increase in bot traffic, we would have detected it.
- The data are consistent with Markdown having a small positive effect, no effect, or even a small negative effect.
- The data are *not* consistent with Markdown being a “game-changer” that doubles bot traffic or produces other dramatic gains.

6. Limitations

6.1 Sample Size

As discussed, the experiment was underpowered for detecting moderate effects. A larger follow-up study is needed to definitively rule out effects in the 10–15% range.

6.2 Metric Scope

This experiment measured crawl volume (bot visits to pages), not citations (inclusion in AI-generated answers). Crawl activity is typically a prerequisite for downstream visibility, but the two are not perfectly correlated. A page could receive more bot visits without appearing in more AI answers, or vice versa. Future work should attempt to measure citation rates directly.

6.3 Duration

The experiment ran for $\sim$21 days. Bot behavior may evolve over longer time horizons, and short-term noise may obscure longer-term trends.

7. Discussion & Implications

7.1 Summary of Findings

We found no statistically significant evidence that serving Markdown to LLM crawlers increases bot visit volume. The CUPED-adjusted mean lift was +13.9% (95% CI: [−11.2%, +46.8%]), but the confidence interval includes zero, and all statistical tests failed to reject the null hypothesis. ChatGPT-User, the dominant bot in our sample, showed a directional advantage for Markdown (+8.5% to +20% depending on specification), but the signal was not statistically significant and was concentrated in high-traffic pages.

7.2 Why No Slam Dunk?

We hypothesize two primary reasons:

1. **LLMs are already proficient at parsing HTML.** These models have been trained on billions of web pages, and the overwhelming majority of the internet is published in HTML [Mueller, 2025, Goodwin, 2026]. Model providers have invested heavily in building robust extraction pipelines that handle messy, nested, JavaScript-heavy HTML at scale.
2. **The internet is still HTML-first.** While Markdown may be cleaner in theory, bot operators are optimizing for the content that exists now—and that content is overwhelmingly HTML. Until a critical mass of high-value content shifts to Markdown, there may be limited incentive for bots to prioritize it.

7.3 Practical Recommendations

Expectation: not a silver bullet. Even in the most optimistic interpretation of our data, any effect is likely marginal and may be confined to pages that already receive high bot traffic.

Focus elsewhere. If you are optimizing for AI visibility, prioritize fundamentals:

- High-quality, crawlable content
- Clear semantic structure (headings, lists, tables)
- Fast load times
- Ensuring bots can access your pages (no aggressive rate-limiting, no bot-hostile CAPTCHAs)

7.4 Future Research Directions

- **Larger sample size:** More pages and more sites would push the MDE below 20%, enabling detection of moderate effects.
- **Citation measurement:** Directly measuring inclusion in AI-generated answers (e.g., via API scraping or user studies) would test the end-to-end visibility hypothesis.
- **Longitudinal tracking:** A 90-day or 180-day experiment would capture longer-term trends and seasonal effects.
- **Content type heterogeneity:** Testing across verticals (technical documentation, news, e-commerce) would assess generalizability.

7.5 Will This Change?

Possibly. LLM crawler behavior is evolving rapidly, and what is true today may not hold in six months. As more sites adopt Markdown, as bots become more sophisticated, and as model providers refine their content preferences, the calculus could shift. We plan to re-run this experiment later in 2026 and will update our guidance accordingly.

8. Conclusion

We conducted a randomized controlled experiment to test whether serving Markdown to LLM crawlers increases bot visit volume. Across 381 pages, six websites, and 20 days, we found no statistically significant effect. While ChatGPT-User showed a directional advantage for Markdown, the signal was not robust enough to inform confident recommendations.

The data suggest that Markdown is not a "game-changer" for AI visibility—at least not yet. Publishers optimizing for LLM crawlers should focus on content quality, crawlability, and structure rather than format [OtterlyAI, 2026]. As bot behavior evolves, we will continue to test and refine our understanding.

A. Bot Classification Rules

Table 13: User agent patterns used for bot classification.

Bot Family	User Agent Patterns
ChatGPT-User	`ChatGPT-User`
OAI-SearchBot	`OAI-SearchBot`
GPTBot	`GPTBot`
ClaudeBot	`ClaudeBot`, `Claude-Web`
Meta/Facebook	`Meta-ExternalAgent`, `FacebookBot`, `facebookexternalhit`
Perplexity	`PerplexityBot`
DuckAssist	`DuckAssistBot`

B. CUPED Regression Coefficients

Table 14: CUPED regression coefficients and variance reduction by specification.

Specification	θ (Overall)	θ (ChatGPT)	Var. Reduction (Overall)	Var. Reduction (Cha
Jan 2 – Jan 12	2.23	2.71	39%	56%

References

Akamai Technologies. (2025, October). *The rise of the LLM AI scrapers: What it means for bot management*. Akamai Technologies. https://www.akamai.com/blog/security/rise-llm-ai-scrapers-bot-management

Cloudflare. (2025a, August). *A deeper look at AI crawlers: Breaking down traffic by purpose and industry*. Cloudflare Blog. https://blog.cloudflare.com/ai-crawler-traffic-by-purpose-and-industry/

Cloudflare. (2025b, October). *The crawl-to-click gap: Cloudflare data on AI bots, training, and referrals*. Cloudflare Blog. https://blog.cloudflare.com/crawlers-click-ai-bots-training/

Deng, A., Xu, Y., Kohavi, R., & Walker, T. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM '13)*, 123–132. https://exp-platform.com/Documents/2013-02-CUPED-ImprovingSensitivityOfControlledExperiments.pdf

Goodwin, D. (2026, February). *Google & Bing don't recommend separate Markdown pages for LLMs*. Search Engine Land. https://searchengineland.com/google-bing-dont-recommend-seperate-markdown-pages-for-llms-468365

Mueller, J. (2025, November 23). *[Bluesky post]*. Bluesky. https://bsky.app/profile/johnmu.com/post/3m6dayx6iyc2r

OtterlyAI. (2026, February). *llms.txt and AI visibility: Results from OtterlyAI's GEO study*. OtterlyAI. https://otterly.ai/blog/the-llms-txt-experiment/

Previsible. (2026, January). *2025 state of AI discovery report: What 1.96 million LLM sessions tell us about the future of search*. Previsible. https://previsible.io/seo-strategy/ai-seo-study-2025/

Search Engine Land. (2025b, August). *AI traffic is up 527%. SEO is being rewritten*. Search Engine Land. https://searchengineland.com/ai-traffic-up-seo-rewritten-459954

Thunderbit. (2026, February). *The state of web crawling in 2026: Key statistics and industry benchmarks*. Thunderbit. https://thunderbit.com/blog/web-crawling-stats-and-industry-benchmarks